CS 5350/6350: Machine Learining Spring 2018

Homework 1

Handed out: 24 January, 2018
Due date: 9 Feburary, 2018

General Instructions

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 10 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by **midnight of the due date**. Please submit the homework on Canvas.
- Some questions are marked **For 6350 students**. Students who are registered for CS 6350 should do these questions. Of course, if you are registered for CS 5350, you are welcome to do the question too, but you will not get any credit for it.

1 Hypothesis Space

Suppose in our supervised learning task, we have 4 boolean features, x_1, x_2, x_3 and x_4; the label y is binary. We have collected a set of training data, listed as follows. To learn a boolean function from the training data, we try several hypothesis spaces.

x_1	x_2	x_3	x_4	y
0	0	1	0	0
0	1	0	0	0
0	0	1	1	1
1	0	0	1	1
0	1	1	0.	0
1	1	0	0	0
0	1	0	1	0

1. [30 points] Conjunctions: each function is a conjunction of a set of variables or their negations. For example, one conjunction function could be $x_1 \wedge x_2 \wedge \neg x_4$. You can choose any set of variables (out of $\{x_1, x_2, x_3, x_4\}$) to create a conjunction function. An empty set of variables is allowed as well.

   (a) [10 points] List all the functions in conjunction space.

   (b) [10 points] List all the functions in conjunction space that are consistent with the training data.

   (c) [5 points] If there are n binary features, rather than 4, what is the size of the conjunction space?

   (d) [5 points] If there are n binary features, rather than 4, what is the size of full hypothesis space that comprise of all boolean functions with n binary variables?

2. [30 points] m-of-n rules: an m-of-n function has n binary variables; the function value is 1 if at least m of these variables are 1.

   (a) [10 points] List all the functions in m-of-n rule space.

   (b) [10 points] List all the functions in m-of-n rule space that are consistent with the training data.

   (c) [5 points] If there are n binary features, rather than 4, what is the size of the m-of-n rule space?

   (d) [5 points] Compared with the conjunction space, is m-of-n rule space more expressive/flexible? Why?

2 Decision Tree

1. [20 points] Decision tree construction.

   (a) [10 points] Use the ID3 algorithm with information gain to learn a decision tree from the training dataset in Section 1. Please list every step in your tree construction, including the data subsets, the attributes, and how you calculate the information gain of each attribute and how you split the dataset according to the selected attribute. Please also give a full structure of the tree. You can manually draw the tree structure, convert the picture into a PDF/EPS/PNG/JPG format and include it in your homework submission; or instead, you can represent the tree with a conjunction of prediction rules as we discussed in the lecture.

   (b) [5 points] Write the boolean function which your decision represent. Please use a table to describe the function — the columns are the input variables and label, i.e., x_1, x_2, x_3, x_4 and y; the rows are different input values and the function values.

   (c) [5 points] As we discussed in Lecture 2 (Supervised Learning: The Setup), when we use the full hypothesis space, it is hard to identify a function in the full space based on a relatively small training dataset. The dataset in Section 1 is an illustrative example: we showed that using the full hypothesis space, we are unable to find out even one function. We have known that the decision tree hypothesis space is as large as the full boolean function space. But now, we are able to use the ID3 algorithm to learn a decision tree, which corresponds to a specific boolean function. Can you explain why? Which steps of the ID3 algorithm enable us to identify a function from such a large hypothesis space?

2. [20 points] Let us review the geometric object classification task mentioned in our class. Please use the dataset shown in the lecture slides — page 16 on Lecture 3: "Decision Tree: Representation".

   (a) [10 points] Use the ID3 algorithm with information gain to learn a decision tree from the training dataset. Note that we only have two attributes in total (color and shape). As in the above problem, please list every step in your tree construction, and the whole tree structure as well.

   (b) [2 points] What is prediction for a new instance "Green Triangle"?

   (c) [8 points] Why can your tree predict the instance? Why can't the example tree predict the instance (see Page 25)?

3. [20 points] Let us use the training dataset for making a decision whether to play tennis or not (Page 41, Lecture 4: Learning Decision Trees). Suppose now, we add one more training instance where Outlook's value is missing: {Outlook: Missing, Temperature: Mild, Humidity: Normal, Wind: Weak, Play: Yes}

   (a) [5 points] Use the most common value in the training data as missing value, and calculate the information gains of the four features.

   (b) [5 points] Use the most common value among the training instances with the same label, namely, their attribute "Play" is "Yes", and calculate the information gains of the four features.

   (c) [10 points] Use the fractional counts as the feature values, and then calculate the information gains of the four features.

4. [**For 6350 students**] [30 points] Please prove that information gain is always non-negative (Hint: use convexity).

3 Programming Assignments

[100 points] We will implement a decision tree learning algorithm for car evaluation task. The dataset is from UCI repository(`https://archive.ics.uci.edu/ml/datasets/car+evaluation`). Please download the processed dataset from Canvas. In this task, we have 6 car attributes, and the label is the evaluation of the car. The attribute and label values are listed in the file "data-desc.txt". The training data are stored in the file "train.csv", consisting of 1000 examples. The test data are stored in "test.csv", and comprise of 728 examples. In both training and testing datasets, attribute values are separated by commas; the file "data-desc.txt" lists the attribute names in each column.

1. [50 points] Implement the ID3 algorithm which supports both the majority error and information gain to select attributes for data splits. Besides, your ID3 should allow users to set the maximum tree depth.

2. [30 points] Use your implemented algorithm to learn decision trees from the training data. Vary the maximum tree depth from 1 to 7 — for each setting, run your algorithm to learn a decision tree, and use the tree to predict both the training and testing examples. Report in a table the average prediction errors on each dataset when you use information gain and majority error heuristics, respectively.

3. [20 points] What can you conclude by comparing the training errors and the testing errors?